# The Unseen Scholars
## *Researching Information in the Digital Age*

*Computers and the World Wide Web have radically changed the way scientists and scholars do their research, including how they gather and exchange information. The Los Alamos Research Library has helped bring about those changes by conceptualizing and developing innovative, web-based research tools. Herbert Van de Sompel, head of the Research Library's Digital Research and Prototyping Team, and Johan Bollen, also of the Prototyping Team, sat down with 1663 to discuss this evolution and to touch upon ways of adapting scientific communication to the digital age.*
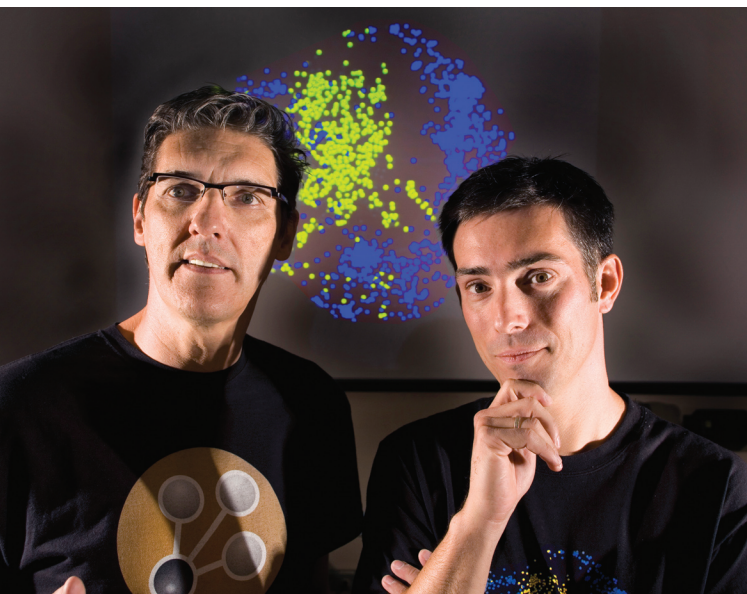


Herbert Van de Sompel (left) and Johan Bollen

**1663:** Many people are surprised to learn that the Research Library employs several Ph.D. researchers, including the two of you. Why does the library need such high-powered talent?

**Herbert:** The library recognizes that our research, which has a significant international impact on technologies and policies for scientific communication, helps to create services that benefit researchers here at Los Alamos.

These days, scientists rarely go to the library to browse the journals for papers; they log onto the library from their computers and use a search engine to discover papers in a collection of thousands of electronic journals. The search engine is a library-provided "service." Other library services let the scientists download papers, extract citations, get alerted to new articles, get recommendations about what they should read next, etc. These services don't emerge out of thin air, and some of the more-advanced ones are built around my team's concepts and tools.

**Johan:** OPPIE, the Research Library's new search engine, is an example. It's very fast, much faster than the previous search engine. OPPIE searches a large archive of over 90 million records, most of them the bibliographic metadata—the title, author(s), abstract, etc.—of published articles. After finding the relevant metadata, OPPIE leads the user to the article itself.

**1663:** And you created OPPIE?

**Herbert:** The Library's Application Development Team created OPPIE. My team's focus was the archive where OPPIE's content is stored. The archive's design is very scalable, meaning we'll be able to store hundreds of millions of digital records without the system crashing or slowing down. The architecture is really novel and was designed and developed by my team. It's very modular, and all components are based on standards, some that I helped develop.

**1663:** What do you mean by standards?

**Herbert:** Standard specifications. Most of my work applies to a level that's way below the computer interface that users see. Basically, I find ways for information systems to work with each other better, and I create specifications that describe how they can do that. For example, a specification might be a set of instructions that tells two servers how to exchange information. Once a specification is released as a standard, it can be adopted by information systems on the Web. There's nothing fancy about it. It's plumbing, like the pipes running beneath the house. People are never aware of the plumbing, but because it's there, they can build fancy bathrooms and kitchens.

**Johan:** Plumbing is Herbert's private joke. Many people are acutely aware of his work because it's had such an influence on the way the academic and research communities access and exchange information.

**1663:** Joke or not, plumbing's a great analogy. Can you give us a concrete example?

**Herbert:** There's the Protocol for Metadata Harvesting (PMH). Soon after the Web emerged, hundreds of scientific pub-

## OPPIE: Bringing the World to the Laboratory

*The Research Library's new, remarkably fast search and discovery engine, OPPIE, was born this year in May. "We worked on bringing this cutting-edge technology into production for about 4 years," says Miriam Blake, director of the Laboratory's Research Library. "The result is a search tool so well designed that it will be able to service the Laboratory's special needs long into the future."*

*The Research Library needs to provide Los Alamos employees with access to the world's scientific information, while preventing the world from knowing what information those employees are seeking. So in 1994, the library began to purchase content—articles and metadata—from publishers and store it in the library's own digital archive. Instead of searching the Web for research papers, Los Alamos scientists search this local archive, and their activities remain confidential and secure.*

*By 2000 the archive had swollen to over 70 million records and was having growing pains. Because of the way data were stored, the archive did not scale, and as the number of records increased, the archive got more difficult to search. Users began to notice that SearchPlus, the search engine that interfaced with the archive, was running more and more slowly.*

*A completely new type of archive, known as aDORe (pronounced "adore") was designed and developed by the library's Prototyping Team. This new archive has a unique architecture that incorporates many of Herbert Van de Sompel's standard technologies (OAI-PMH, OpenURL, OAI-ORE, info URI) and is highly scalable.*

*The library's Application Development Team then built OPPIE on top of aDORe, converting the millions of records into a standardized format and loading them into the distributed aDORe archive. The team built the OPPIE interface that researchers use to search the archive and continues to add other tools, many of which are implemented using freely available open-source software. OPPIE can run without expensive commercial products and should be well supported by the open-source community.*

*"We used widely accepted standards and open-source tools to make OPPIE sustainable and compatible with other systems. There is immense flexibility," says Blake. "We can plug in new tools and features very easily, so as the Laboratory moves into the future, we can be very responsive to evolving customer needs."*

**OPPIE** science starts here

lishers around the world started making their journals and associated article metadata available online. That was a good thing. The bad thing was that one had to search each publisher's metadata separately. In order to overcome this problem, you wanted to collect the metadata into one large pool and search it there. But there was no uniform way to collect metadata from the publishers' information systems, and they used different metadata conventions.

**Johan:** It was crazy. You couldn't just tell a search engine to look for an author; you had to do multiple searches in several systems. But there were hundreds of publishers, and you couldn't cover all the bases.

**Herbert:** In 1999, Paul Ginsparg, who created the Los Alamos preprint archive, Rick Luce, then the director of the Research Library, and I founded Open Archives Initiative (OAI). Its goal was (and still is) improving the dissemination of scholarly information through technical means. Under the OAI umbrella, several colleagues and I began to develop a protocol, a set of commands that would tell one computer system how to present metadata in a standard way, no matter how it was stored internally, so another computer system could grab it. The protocol created an interface for metadata exchange between the two systems, and it became a standard.
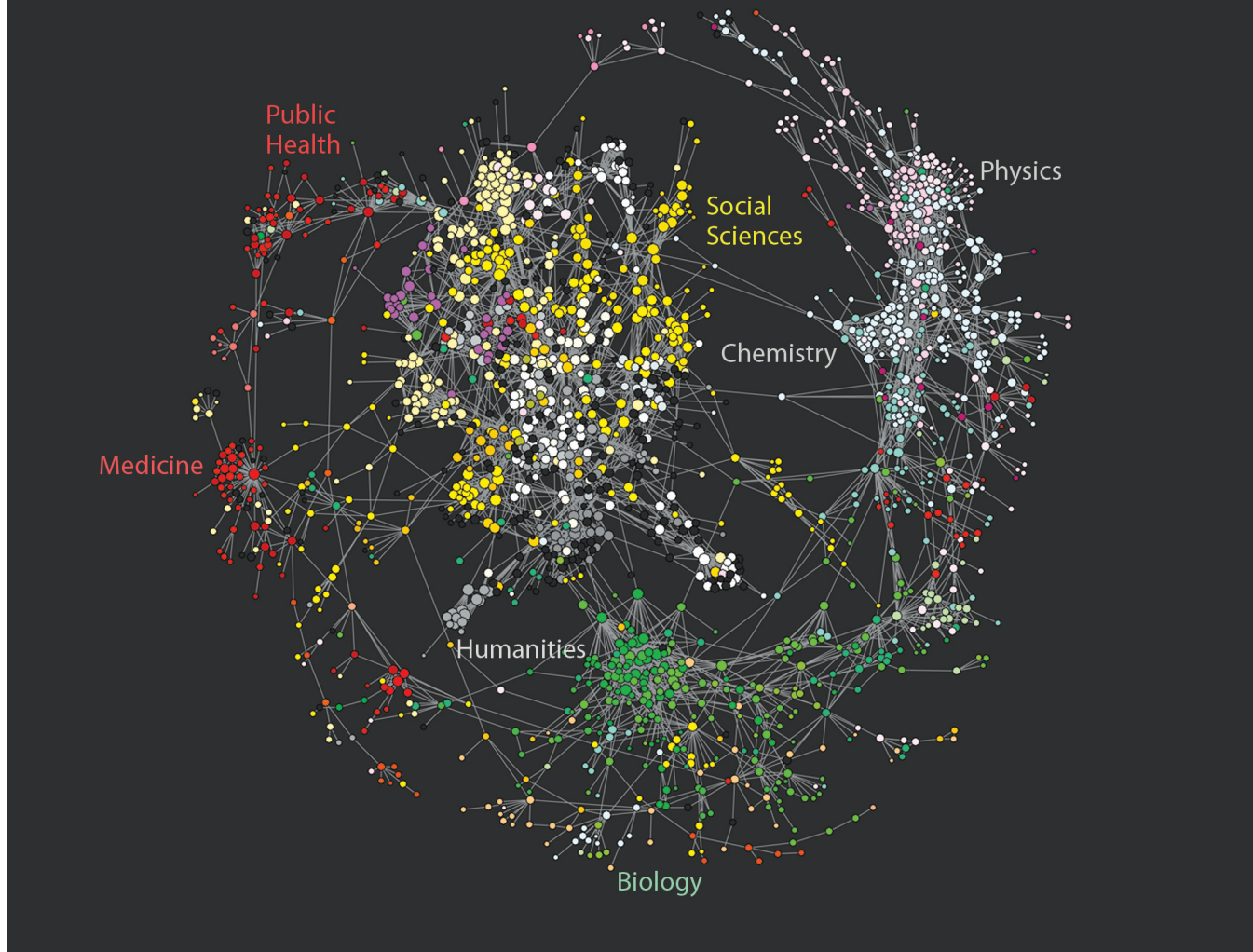
Systems around the world now use PMH in a variety of ways. Our own OPPIE uses it to obtain its metadata from its underlying content archive. Via PMH, OPPIE checks whether new content is available and if so, grabs it (also via PMH) and adds it to the search engine. OPPIE harvests about 90,000 new records a week in this way.

**Johan:** But once a scientist searches for and finds a paper, he or she wants to read it. Just a few years ago, there were some more problems involved in doing that, but another of Herbert's ideas, SFX, fixed it.

**Herbert:** The situation was this: a scientist logged onto the library to read a research paper and found in it a link to another paper. The first paper was in a journal published by Elsevier, a large publishing firm, while the second was in a journal published by Wiley. Because the library has a subscription to access content from both publishers, the scientist should have been able to click on the link and read the second paper, but when he clicked on the link, he was told, "Sorry, you don't have a license to access this article."

The problem was that many institutions would subscribe to a publisher indirectly through a content aggregator like

Above: In this "map" of real usage data, each dot is an electronic journal; a line between dots indicates that people accessing one journal went on to access the other. Shorter lines mean a stronger correlation. The map shows an interconnected cluster of journals from the social sciences and humanities surrounded by a network of journals from the natural sciences. The flow of information between the two domains is largely through interdisciplinary fields.

Ebsco, which provides access to lots of electronic journals. Elsevier, however, would send everybody to Wiley indiscriminately, even though the scientist should go to Ebsco, because that is where their institutional subscription existed. It was a huge problem.

**1663:** And the fix?

**Herbert:** The fix was SFX, a link server, or a computer that acts as a sort of concierge. Once installed at an institution, it knows all about the institution's subscriptions and services. Now when a scientist clicks, Elsevier's link goes to SFX at the scientist's institution because that link server knows which publisher or provider the scientist should get the Wiley content from. SFX sends the scientist a pop-up message that says, "Click here to access this paper."

**Johan:** But Herbert didn't tell every institution around the world to buy an SFX link server. Instead he created a standard—OpenURL—that specifies how an information system such as Elsevier should link to a link server such as SFX. As a result, many commercial link servers were developed, and most academic institutions worldwide now use one. OpenURL is even supported by Google Scholar.

**1663:** That's very clever.

**Johan:** It gets better. The network of link servers opened up a whole new area of research. Every user who clicks on a link is announcing to his institution's link server, "I want to access this now." So the link server can maintain a log file of what's being accessed and when. The log is called usage data.

Herbert and I realized that if we could access the usage data of researchers worldwide, we could build an incredible picture of what is going on in science. The Andrew W. Mellon Foundation, a philanthropic foundation interested in scholarship and new tools for scholarship, came to the same conclusion and funded the MESUR project, which I've been working on for the past few years. One goal is to see if usage data will give us a way to assess scholarly impact, that is, to see who are the most-influential people, which are the most-important journals, what are the critical institutions, etc.

The value of a research paper is currently assessed using citation data. People literally count how many times the paper is cited. It's assumed that good papers get cited more often. But citation data provide a view of how science existed several years ago. It may take a year before a scientist's idea

is written up, peer reviewed, and published, and then an equally long time before another scientist reads the paper and writes a new paper that cites it. It often takes several years for citations to mature in a particular discipline. If the number of citations is the only metric used to assess scholarly worth, young researchers who have been publishing for only a few years may be undervalued.

**Herbert:** And there is more that citation data do not reveal. Say a journal doesn't get cited much but is read by both physicists and archeologists. The journal fosters the flow of ideas between the two fields. Citation data do not reveal this because a physicist will typically not cite the archeology paper, but usage data show the connection.

**Johan:** We've collected perhaps the largest existing set of usage data in the world—over a billion "clicks" gathered over the years from some of the world's most-significant publishers and aggregators and a set of institutional consortia that includes the University of California, California State University, the University of Texas, and lots of others. It's an enormous dataset that we believe covers a good chunk of the online activity pertaining to research in science and the humanities, including medicine.

**1663:** Did you have to twist arms to get institutions to part with their usage data?

**Johan:** I often joke that all of my gray hair has been acquired in the past year from begging these people for their usage data. Really, most were eager to collaborate with us, in part because of the reputation that our team and the Research Library have in the community. They also know the data have value; they just don't know how to exploit the data yet. I tell them right off that usage data can be used to assess value because they reveal immediately how many people are reading which papers. That information could be used, for example, to price the journals or to reward the authors. And the value assessment would be statistically more accurate than a citation-based value because a poorly cited paper may nonetheless be read thousands of times.

But as Herbert said, we can also look at relationships between papers, or between journals, and define, say, a "bridge value" metric that quantifies to what extent a paper connects normally disparate groups. We've come up with dozens of metrics that can be used to measure value and to improve our understanding of science.

**1663:** Wow! You may change the entire notion of what constitutes a good research institution or who should get tenure.

**Herbert:** That's a general theme of the Prototyping Team's work: use the new capabilities of the digital era to improve scientific communication. Another example is the Object Reuse and Exchange project (ORE), which we worked on for the past two years. Its starting point was the consideration that in so-called eScience, a publication is not just a paper, but rather the aggregation of a paper, a dataset, maybe a video recording of a computer simulation, some software, etc. All these resources sit on different Web servers, but they form a logical whole—a digital-era scientific publication. So, somehow we must be able to express that these distributed resources belong together. We need to glue them together.

The Web gives us a fantastic mechanism, the URI, to talk about each of those resources individually by means of its Web address. It does not give us a way to talk about an aggregation of resources. I have worked with my team and with colleagues around the world to give the Web the ability to handle such aggregations. The resulting solution is based on the principles of the Semantic Web—the Web for machines—and the specifications were recently published. The Mellon Foundation, the National Science Foundation, and Microsoft funded this project. There are already groups in the United States, Europe, and Australia implementing these new specifications, and also the library is developing compliant tools. Pretty cool.

**1663:** Scientific communication will never be the same.

**Herbert:** Not if we have it our way. ❖

—*Jay Schecker*

Above: The open-source djatoka (with silent "d") image server, designed and developed by Ryan Chute, is another of the Prototyping Team's realizations. It facilitates delivering high-resolution images, such as the ribosome simulation shown here, to a Web browser. Just two months after djatoka's release, both scientific and cultural heritage institutions have already expressed a strong interest in the software.